Policy on the AI Exponential

Dario Amodei

Contents

- 1. Regulation and public safety
- 2. Macroeconomics and tax policy
- 3. Accelerating AI's positive impact
- 4. The state and civil liberties
- 5. Securing leadership by democracies
- A window of opportunity

June 10, 2026

In one of the side plots to *The Lord of the Rings*, two of the Hobbits attempt to rouse Treebeard—a wise but ponderous sentient tree—to defend his forest from an army that is cutting it down. The problem is that Treebeard operates at a very different speed than the Hobbits. It takes him a full day simply to say hello to another tree, so getting him and his peers to act fast enough is nearly impossible.

The intersection of AI and our political institutions feels a bit like the Hobbits and Treebeard. AI is advancing at a lightning pace—in only four years, AI models have gone from barely being able to write a coherent line of code to writing most of the code at major AI companies. Similar gains have been made in biology, physics, math, finance, law, translation, and many other fields. AI's scaling laws, which predict an exponential increase in general cognitive capabilities with increasing computing power, now have over a decade of empirical evidence behind them. If these scaling laws continue for only a year or two longer, we are likely to get what I've called *Powerful AI,* or "a country of geniuses in a datacenter".

By contrast, policy—and especially legislation—moves very slowly. Often this is for good reasons: governments have grave powers, and it's usually for the best that they aren't used too hastily. But the mismatch in timescale is nevertheless very painful: in the several years that it can take Congress to act, AI can go from an amusing toy to the full country of geniuses.

Over the last few years since AI has become a major commercial technology, those of us who wanted to handle it responsibly have faced a dilemma. We could see clearly where the exponential was going: we strongly suspected that within a few years AI would be one of the rare technologies that fundamentally reshapes the entire policy landscape, in the same way that nuclear weapons reshaped geopolitics and the industrial revolution fundamentally reshaped every economic and social issue. But to those looking only at what AI could do *at the time*, it looked like a much more mundane technology—similar perhaps to the latest consumer app or cryptocurrency. It was hard to convince most policymakers and companies that anything other than a *laissez faire* attitude made sense. And to be fair, the fact that AI's radical effects were not yet present, and that we didn't know exactly what shape they might take, made it difficult to design the right policies even if there had been the will to act.

Given the limits imposed by this situation, many safety advocates (including Anthropic) have so far been focused on advocating for policy actions that preserve optionality, tee up a fast reaction in the future, or give the world better insight into what is coming down the pike – things like transparency legislation, export controls on chips, and data collection on AI's labor effects. These are not enough, but they have felt like all that was possible.

In the last few months, however, the evidence of AI's incredible power, as well as its risks, has become undeniable. Perhaps the most emblematic example is Claude Mythos Preview and the discovery that frontier models pose very real risks to cybersecurity, creating the potential for disruption of the financial sector, critical infrastructure, and national security. Mythos Preview scrambled the global cybersecurity landscape. But its broader significance is that it proves beyond doubt that AI models are now tools of global and national strategic consequence. The cyber risks that Mythos-class models present will not be the last that we must face. I

believe that biological risks may soon follow, and that serious AI autonomy risks may not be far behind[1]

We now, globally and collectively, need to activate a slow and rickety policy apparatus to deal with risks and opportunities that are going to compound surprisingly quickly from here. Many policymakers are showing increased openness to taking action, and it's been encouraging to see our peers come around to the same positions we've been advocating for over the past few years. This is good, but I worry that these early actions are at least a year out of step with AI's rapid progress. This essay is an attempt to close that gap: to lay out where the exponential is now, and the collective action needed to meet the moment.

I will focus on five perennial policy areas that need re-imagining in an AI world: regulation and public safety, macroeconomics and tax policy, scientific innovation, the balance of power between state and society, and geopolitics. I will speak primarily in terms of US policy since Anthropic is an American company, but most of my recommendations are also relevant to the rest of the world.

Along with this essay, Anthropic is releasing a legislative proposal on frontier model testing and a policy framework for job displacement, for which we intend to provide substantial financial backing. We plan to do much more in the future, but we view these as first steps to signal our seriousness.

1. Regulation and public safety

Every new technology or product has both beneficial and harmful uses, and therefore presents a dilemma between innovation and safety. Regulating products makes them less likely to cause harm and has played an important role in improving lives around the world, but it can also directly reduce their benefits and indirectly disincentivize innovation. There is also the Hayekian point that regulators often lack the information needed to make the right decisions about complicated economic tradeoffs, so that regulation is often both ineffective *and* burdensome. A related idea is the Collingridge dilemma,

which states that the impacts of a technology are often hard to anticipate until it is too late to easily manage them.

These dynamics loomed large for AI in 2023-2024. It was clear to Anthropic that AI *might* in the future be capable of producing biological weapons that could threaten millions, or autonomous misbehavior that in extreme cases could even threaten humanity itself. Less clear was the exact *form* in which the risks would appear, how best to test for them and mitigate them, and how they would play out in practice. There was therefore a high risk that legislation written ahead of time would end up being ineffective—creating pointless or low-value compliance requirements while missing the most crucial sources of actual risk[2]

.
Ultimately, we concluded that the right approach at that time was *transparency*. Developers of AI models should have to *disclose* their safety procedures and the tests that they run on their models and report on any critical safety incidents, so that the public and the scientific community could gain better visibility into risks as they emerge. When and if risks become more definite and their shape is more clear, then the evidence gained through transparency could be used to design smart legislation to precisely target the most concerning risks. Thus, in 2025, Anthropic supported transparency legislation, helping to pass SB 53 in California, RAISE in NY, SB 315 in Illinois (in early 2026), and advocating for a transparency standard at the federal level.

However, now the risks are clearly here. It is time to go beyond transparency to more serious and binding regulation of AI. I believe the best analogy, at least at the current stage of the exponential, is to cars, airplanes, or drugs—powerful technologies essential to the modern economy, but capable of killing large numbers of people if designed or operated poorly. I therefore believe we should model AI regulation on agencies like the Federal Aviation Administration (FAA). **Frontier AI models, like airplanes, should be required to go through technical testing and auditing, and their release should be blocked or reversed as a threat to public safety if they do not meet high standards of safety.** I am grateful to see the Trump administration's Executive Order move incrementally towards a

greater role for government in AI, though Anthropic's proposal recommends even further action. Our proposal includes the following elements:

- Models above a threshold of compute should undergo mandatory testing by a qualified third party for their level of risk in four specific areas: cybersecurity, biological weapons, loss of control of AI systems, and automated R&D that could accelerate these other risks.
- The government should have the power to block or deter deployment of the model if it is determined, in light of third-party assessment, to present unacceptable risks. This power must be scoped to the above four specific risks and there must be protective measures against political favoritism or arbitrary decisions.
- Third-party evaluation could be done by a government agency (similar to the FAA) or a set of private organizations that are authorized and inspected by the government to evaluate models according to certain standards (a "regulatory markets" approach).
- AI companies that develop advanced AI models must have strong security standards that protect their model weights, should conduct regular red teaming and penetration testing, and should work with the government to defend against major threat actors.
- Safety incidents in the four critical areas must be reported promptly.

There may come a time, perhaps relatively soon, when we need to go beyond this, when the most powerful AI systems look less like airplanes or automobiles and more like weaponizable nuclear materials—a threat to humanity rather than "just" a threat to public safety. If that occurs, we may need more aggressive regulatory measures than those I have laid out[3]

But just as it was difficult in 2024 to target and apply the measures I'm suggesting now, I don't think we should get ahead of ourselves. We should design policies for the dangers that are emerging today, while laying the foundations to ramp up our response even more quickly as new dangers appear.

2. Macroeconomics and tax policy

Governments have long faced the problem of how to encourage economic growth while also providing important public services and ensuring that the least fortunate are taken care of. An important (and generally correct) premise of these debates has been that *economic growth is fragile and difficult to achieve*—that while reducing inequality might provide important benefits, it has to be traded off against the economic drag of increased taxes or deficits.

I suspect that powerful AI may scramble this assumption. If AI achieves the ability to do most cognitive tasks far better than humans, it stands to reason that it could result in extremely rapid and robust economic growth via the acceleration of science, technology, and operational efficiency. The iterative ability of AI to build even better AI may supercharge that growth even further. But for exactly the same reasons, AI may also act as a more general economic substitute for human cognitive abilities than previous technologies have, while also altering the economy far faster than previous technologies have. Thus, it's reasonable to think that AI could produce much larger disruptions to the labor market than previous technologies, and, potentially, more *enduring* disruptions. We risk ending up in a world where the economic tradeoff dial is stuck on the hypergrowth, hyper-inequality setting, and is potentially very hard to unstick from that setting. *The key challenge in such a world won't be incentivizing growth, but finding a way for everyone to share in the benefits*.

Of the topics discussed in this essay, macroeconomics and enduring labor displacement are arguably the ones that have attracted the most public attention and the most misunderstanding, so I want to be extremely clear on two points.

First, enduring job displacement is undesirable and dangerous, and we should do everything we can to minimize or prevent it, not to bring it about. I have warned about job displacement in interviews and essays because I want both policymakers and the private sector to have the best chance to adapt and respond, not because I am trying to be a "prophet of doom". As a company, Anthropic always does as much as it can to work with customers to find creative new use cases and new

sources of revenue that allow them to do more with their existing workforce, rather than focusing solely on cost savings (which often means reducing the workforce). We also constantly try to think of new interaction paradigms that allow humans to have as active a role as possible in collaborating with AI systems as those systems advance. More broadly, it is valuable for the whole world to experiment with using AI in as many new ways as possible, as that is the way for society to discover new possible job configurations. I do think AI will enable a number of new economic opportunities. I've predicted that AI will enable single individuals to create billion-dollar companies, and we're already seeing teams of only a few people build businesses with hundreds of millions in revenue. But at the same time we should recognize that there's a decent possibility that, despite all our efforts, AI still causes significant enduring job loss—and that this may be an *intrinsic* property of the technology and the way it broadly replicates human cognition[4]

Second, any response to AI-driven job displacement needs to address *both* the need to provide for everyone economically, *and* the need for people to find meaning, purpose, and agency. The latter is ultimately more important, and it depends on deep questions about how society is organized, what people should strive for, and what constitutes the good life. I am actually very optimistic that, even in a world with AI's that are better than everyone at everything, humans can live lives of deep purpose and strive to build awe-inspiring and beautiful things[5]

But this is something to be collectively worked out by society as a whole, not something policy can directly address. Policy can be most helpful in buying us time to do that work, by slowing down job loss and providing economically for those likely to be affected.
In that spirit, some key policy interventions that are likely to be helpful include:

- **Measurement and tracking.** It's easy to dismiss mere data collection and analysis as inadequate to the scale of the problem, but we are unlikely to get good policy if we cannot accurately measure what is happening on the ground. Anthropic has been operating an Economic Index of how people use Claude for nearly a year and a half, but governments have access to types of data we do not, and could greatly expand their economic statistics to more carefully track AI job displacement.

- **Pro-employment incentives.** A wide range of pro-employment policy incentives can help to slow or reduce job displacement, including: wage insurance policies that compensate people when they have to take a lower-paying job[6]

  , retention tax incentives to encourage employers not to make layoffs, workforce training grants, or infrastructure to facilitate matching of employers to employees to speed the rate of labor market adaptation. While the particulars of which interventions are best will depend on what kind of labor displacement AI brings, we should readily accept the costs and market inefficiencies that these policies could entail, particularly as they are likely to be offset by AI-driven productivity gains.

- **Long-term macroeconomic support.** If AI-driven labor displacement ends up being large in magnitude and permanently drives down the demand for labor, it will likely be necessary to go beyond mere incentive programs to long-term income support for a significant fraction of the labor force. Mechanisms such as universal basic income could be financed through taxes on relevant companies or raising the capital gains tax. Universal capital accounts offer another vehicle. Broadly speaking, fast economic growth should create the tax base for shared prosperity.

A common focus of economic concern about AI that I haven't mentioned has been datacenters and particularly their potential to raise energy prices. My view is that AI companies should pay to absorb rate increases—and Anthropic has already made a pledge to do so—but I see public hostility to datacenters as largely a symbol or outlet for broader economic anxieties about AI. It is important we have a direct societal conversation about these wider economic issues and truly have compelling solutions for them, or else they are likely to manifest indirectly, as they have with datacenters.

3. Accelerating AI's positive impact

Just as we must grapple with the balance between innovation and safety for AI itself, we must grapple with the same balance for technologies that are likely to be accelerated by AI, such as biomedicine, energy, or materials science. But while AI itself is likely to present novel challenges that emerge very quickly and that we have no prior experience in handling, other fields accelerated by AI are likely to encounter a very different problem: regulatory systems that were designed for a slower pace of innovation and are not prepared to handle the deluge of new products and advances that AI will bring. AI may also make these downstream technologies safer and more predictable in a way that violates the skeptical assumptions of regulatory agencies like the Food and Drug Administration (FDA).

Thus, for downstream applications of AI—in contrast to AI itself—I am more worried about the regulatory apparatus *slowing down* progress (because it can't handle the increased pace of change) than I am about it failing to address important risks. The last thing we want is for the benefits of AI to be slowed while its risks loom large, so it's important to take action on this problem as soon as possible.

The problem and its solutions will manifest differently in each area of science, commerce, and technology, so I'll focus on one illustrative area: biomedical innovation. This is both because it will likely be the source of AI's biggest humanitarian benefits and because it is an area where regulation is especially complex. We don't know exactly how AI will accelerate biomedical innovation, but it seems likely to:

- Greatly increase the rate at which new drug candidates enter the regulatory pipeline;
- Increase the effect sizes and improve the safety profiles of new drugs, because of better optimization and perhaps better understanding of their underlying biology;

- Develop drug candidates for diseases that have never been successfully treated before;
- Rapidly create entire new forms of therapies, similar to how antibodies, peptides, and cell therapies have become new categories of treatment over the last few decades.

Some of these advances will naturally accelerate regulatory timelines without need for structural change. Drugs with larger effect sizes can lead to smaller, less expensive clinical trials, and activate mechanisms for accelerated approval. But the regulatory system is currently designed to apply a high level of scrutiny and many stages of testing, under the assumption that drug candidates often don't work and often have serious safety problems even when they do. With both the FDA and the European Medicines Agency (EMA), the typical time for a drug candidate to pass through the regulatory pipeline is 7-8 years, in part due to these pessimistic assumptions. Without reforms, AI will simply jam or overload this system.

Obviously, we don't want to change things in a way that leads to a crop of snake-oil drugs or widespread safety incidents. But some relatively simple reforms could make the FDA, EMA, and similar agencies more adaptable to a rapid AI-driven scientific acceleration if one were to occur.

Many of the steps in the clinical process that previously required expensive and slow experiments may soon be done via AI simulation or analysis. Regulatory agencies should consider developing standards *now* for what it would take to accept such methods. This would mean they can be adopted quickly once they work, rather than there being an extended period during which unnecessary tests continue to be required. Areas where this could apply include:

- AI-based pharmacodynamics and pharmacokinetics (PD/PK) modeling;
- Prediction of toxicology to avoid the need for multiple species animal toxicology;
- More accurate dose selection, to reduce to the need for large dose ranges in trials;
- Biomarker validation via analysis of large datasets;

- Synthetic control arms in clinical trials, to reduce the need to recruit more participants;
- Developing surrogate endpoints (particularly important in aging and neurodegeneration).

Beyond these specific examples, agencies should also consider more radical and flexible mechanisms for accelerated approval. If my predictions about AI are correct, there will soon be many instances of interventions that work really well out of the blue, and the regulatory system should be prepared to take them seriously and not adopt a posture of excessive skepticism.

Biomedical acceleration should substantially increase AI's benefits, but it's worth noting that it may also help to reduce AI's risks. Reforming biomedical approvals may help with biodefense, and AI-driven biomedical progress may also improve mental health, which could have a stabilizing effect on society.

4. The state and civil liberties

Every system of government has to confront the question of the state's power and its limits. The state has a legitimate, often existential, interest in protecting its population from internal and external threats. But granting it too much power is the road to tyranny. Modern democracies have largely managed this balance successfully, but it is an uneasy one at the best of times. Enforcing it has required a great deal of legal and constitutional machinery built up over centuries—for example in the United States the First, Fourth, and Fifth amendments, the Posse Comitatus Act, FISA, and so on.

AI threatens to upset this balance while also dramatically raising its stakes. But if we react quickly and meet the moment, we can use AI to create a world that has more robust and durable guarantees of liberty *and* better defense against threats, than we have ever had before.

Powerful AI in the wrong hands could be the ultimate tool of autocracy, and our existing legal and constitutional protections are not fully equipped to counter this threat. Fundamentally, the enormous returns to intelligence in terms of power in the world, combined with the rapid pace of AI's progress, creates a perfect storm for a surprise seizure of power by a range of dangerous actors[7]

The danger could take a variety of specific technological or operational forms, but what they all have in common is the idea that AI could suddenly confer enormous power while routing around existing mechanisms of democratic oversight. A fully automated drone army that sounds like science fiction today could, in the future, obey unlawful orders and allow governments to unilaterally entrench their power; professionally-trained humans are more likely to object to such illegal direction. A surveillance-focused AI could analyze widely available information at massive scale and use it to infer the innermost details of every citizen's life—a technological ability not contemplated by current civil liberties law. All of this could happen very quickly, or in secret, so it is important to proactively fortify democracies' commitment to freedom and civil liberties.

The following are some policy ideas we should consider:

- **Create reliable accountability rules for fully autonomous weapons.** Autonomous weapons, and especially any autonomous systems that coordinate or direct them, should be required to respond to mechanisms of constitutional and command accountability (e.g. court orders, legislation, and accountability to senior human overseers) rather than blindly following orders. This could mean that a suitably-designed legal review panel or the judicial branch have their finger on an "off switch", that the systems themselves are intrinsically trained to seek out and respond to legitimate oversight authority, or both.

- **Ban the domestic use of fully autonomous weapons.** While there is a legitimate case for the necessity of fully autonomous weapons to defend against foreign adversaries (such as Russia invading Ukraine), there is no justification for their use against Americans. The military already has some

limits on its ability to operate domestically, but ideally these weapons should be banned in law enforcement as well.

- **Close the bulk collection / data broker loophole.** Under current law, data that Americans share with private companies (such as internet providers) can be purchased and used for bulk analysis in domestic surveillance and law enforcement. This gap in privacy protections predates AI, but AI will raise the stakes considerably by making mass analysis of such data far more revealing and useful than it has been in the past. This loophole should be closed.
- **Public rights to AI advice during adverse government action.** As a general principle, it seems important that any person or organization that is the subject of adverse government action (e.g. regulatory or legal action) has access to AI that is at least as capable as whatever the government is allowed to use in that particular action. This would mean not giving the government an unfair advantage, effectively undermining citizens' legal rights. This could be added as an extension or interpretation of the Administrative Procedure Act, due process protections, or the Sixth Amendment right to legal representation.

Finally, it is worth noting that governments are not the only entities that we should beware of when it comes to AI-driven seizure of power. At various times in history (such as the Gilded Age in the United States or the East India Company in the UK), companies have become powerful enough that they capture the state or adopt quasi-state characteristics. AI will soon become so capable that I worry it cannot safely be fully entrusted to *either* governments or companies, and there must be checks and balances on each.

Regulation is one answer on how to rein in companies (and my ideas for that are in Section 1), but it's also important that AI companies have more separation of power and accountability than is typical for private entities. Anthropic's Long-Term Benefit Trust (an independent governance body designed to hold the company to its mission) is one such structure, and the industry should continue to explore mechanisms that go further. Getting the balance right—so that both companies and the government have meaningful checks on their powers—is essential.

5. Securing leadership by democracies

It has become a common instinct, perhaps developed from recent experience with the internet and telecommunications, to regard new technologies geopolitically as instruments of trade policy, with the aim being to "diffuse our technology stack around the world". But it is my very strong belief that AI is something much more profound, something that resets the whole game board and around which all future geopolitical strategy must be shaped—like nuclear weapons, but potentially even more so.

If AI really will soon be "a country of geniuses in a datacenter", or anything remotely close to it, then **AI is likely to be the dominant source of military and economic power for any nation**. In a virtual country of 100 million geniuses, 10 million could be applied to military strategy, 10 million to drone manufacture, 10 million to weapons R&D, 10 million to intelligence collection and analysis, 10 million to general scientific advancement, and so on. A nation that possesses powerful AI facing one without it—or even facing one that is behind in AI by 3 years—could be the equivalent of an army of World War II Marines facing an army of medieval swordsmen.

In addition, if powerful AI enables deeper and potentially permanent forms of autocratic repression (see Section 4), this makes it all the more important that the world's most powerful nations are democracies—or at least that strong protections exist against AI-driven repression. It also increases the urgency of a focused geopolitical strategy.

Democracies should seek to form a global coalition centered on building AI according to their common values, iteratively trying to draw in the rest of the world by making it more and more attractive to be part of the coalition and less and less attractive to be outside it. The coalition should be a coordinated internationalization of the AI policy ideas discussed in Section 1 through 4, plus an

effort to lock down the supply chain critical to building AI by sharing it within the coalition and denying it to those outside it. Some principles and operating goals might include:

- **Managing the AI supply chain.** Members of the trusted coalition should freely share chips and semiconductor manufacturing equipment (SME) with each other, while working together to deny it to adversaries. US export controls on frontier chips and SME to China have been a major contributor to the US's overall lead in AI, and these policies need to be expanded, tightened, and coordinated with other likeminded states. Pending legislation like MATCH and OVERWATCH is a good first step here, and allied democracies need to consider similar measures.
- **Coordinate to address AI's risks.** The policies to address biological, cybersecurity, and autonomy risks described in Section 1 will be more effective (as well as less burdensome to industry) if they are coordinated internationally. This would mean companies can comply with compatible standards and regulators can learn from each other how to best measure and mitigate these risks. Law enforcement and intelligence agencies should also work more closely together on tracking and disrupting threats of misuse, such as efforts by terrorists to build biological weapons with AI.
- **Share AI's benefits.** Trade and regulatory policy can be used to facilitate a more rapid diffusion of AI's economic benefits within the coalition, sharing lessons on how to accelerate innovation. Coordinating approaches to beneficial deployment could help bring the benefits of AI to developing countries. For example, harmonization of medical approval regimes could lead to faster and better testing and approval of AI-enabled drugs (as discussed in Section 3 above).
- **Mutual defense.** Countries in the coalition should work together to defend each other with AI and from adversaries' AI. The coalition should collectively ensure sufficient production of AI-led cyberdefenses, AI-powered drones, AI-driven manufacturing, classified AI compute, AI-driven R&D, and sharing of AI-driven intelligence collection.
- **Rejection of AI-powered repression.** Coalition members should have to reject the high-tech, ultra-repressive, AI-powered tyranny that I warned

about in *The Adolescence of Technology*, and must have safeguards similar to those I described in Section 4 above.

- **Macroeconomic cooperation.** Crises of employment or job stability, like any other economic crisis, can be contagious across borders. Countries therefore have a mutual interest in working together to coordinate macroeconomic support and stabilization policies, like those described in Section 2, to counter any employment effects.

The goal should be to make membership in the coalition as attractive as possible—and the costs of remaining outside it clear. The coalition would rest on coordination among sovereign states, with each nation retaining full authority over its own affairs. It could grow iteratively, starting with ideologically aligned democracies (which will be naturally amenable to joining) and progressively welcoming countries that are less naturally aligned but prepared to meet the coalition's standards in exchange for the enormous benefits of membership. Ideally, the entire world would eventually join. But even if that isn't possible, building the coalition puts democracies in the strongest position to contain and outcompete the regimes that remain committed to repression.

A window of opportunity

AI's exponential progress has created an urgency and a pace of change that the policymaking process is ordinarily ill-equipped to handle. But it has also created a unique window of opportunity. The confluence of clear and present evidence of AI's risks, an early taste of the AI's potential for both economic value creation and economic disruption, and a remarkable public backlash against unregulated approaches to AI have created a situation where policymakers are unusually open to forward-looking actions. Treebeard and his forest are waking up.

It's become popular in AI industry circles to view this as a PR problem: to say that AI needs "better marketing". I reject this framing completely. People are worried about AI because they *correctly* perceive that its risks are real, not because AI CEOs

have been insufficiently Panglossian. I believe it is my duty as an AI leader to continue to be transparent about these risks, and public concern in response to this transparency constitutes democratic accountability working as it should. The key challenge is focusing this concern into constructive solutions and not allowing it to descend into formless anger and violence.

I am optimistic about finding solutions because many of these issues—from addressing job displacement, to pre-release testing of models, to export controls on chips, to other AI related policy issues such as energy use—have a common-sense appeal across the political spectrum. There is an aspirational but realistic future world in which a broad nonpartisan coalition, driven by direct recognition of the challenges posed by AI, leads to sane and forward-looking policies being adopted much faster than usual. The sooner we do this, the sooner we can all share in AI's incredible benefits.

I would like to thank Allan Dafoe, Mariano-Florentino Cuéllar, Richard Fontaine, Buddy Shah, Vas Narasimhan, Matt Yglesias, Nick Beckstead, Jason Matheny, Brad Carson and many of the staff at Anthropic for their comments and feedback on drafts of this essay.

Footnotes

1. [1] I discuss biological and autonomy risks, among others, in my essay *The Adolescence of Technology*. The Anthropic Institute has also released some initial internal data in *When AI Builds Itself* about the possibility of recursive self-improvement, or models that are autonomously capable of building better models.↩

2. [2] This phenomenon is not theoretical: we've observed it multiple times in our own voluntary governance frameworks like our Responsible Scaling Policy. If we give ourselves a fixed or rigid list of safety requirements for

future AI models, a very likely outcome is that requirements which turn out to matter very little end up consuming 95% of our compliance efforts, while at the same time we discover that some of the biggest sources of risk weren't anticipated in our list at all. Voluntary frameworks can be changed and adapted, but this is much harder with legislation. My attempts to wrestle with this dilemma can be seen in my two public letters about SB 1047, a 2024 California law that attempted to address catastrophic risks and about which I had mixed feelings for the reasons above.↩

3. [3] For example, truly severe biological risks may be much more difficult to manage than cyber risks, because attackers have a strong advantage relative to defenders and the severity of a catastrophe may be much greater.↩

4. [4] See *The Adolescence of Technology* for a more detailed analysis why the logic that has led to rapid job market recovery and a lack of enduring labor displacement in other technologies may not apply to AI, and in particular why the usual adaptive mechanisms like Jevon's paradox or comparative advantage may be overwhelmed by the pace of the technology.↩

5. [5] As an example, people still devote their lives to playing Chess, or Go, or climbing mountains, and are still revered for these activities, even though all can be done better by machines.↩

6. [6] This essentially gives people an extra incentive to migrate to a new job and start training for a new career ladder, even when it may be painful in the short run, by paying them the difference between their new and old salaries, if the new one is lower.↩

7. [7] See *The Adolescence of Technology* for more on this topic.↩

Back to top

Privacy policy